

Using Apollo at the i5k Workspace@NAL

Monica Poelchau, USDA-ARS NAL

October 29th, 2019



Agenda

- Manual annotation general overview
- 15k Workspace tools for manual annotation
 - BLAST, Clustal, HMMER
 - Apollo2
- Manual annotation example: preparation
- Manual annotation live example

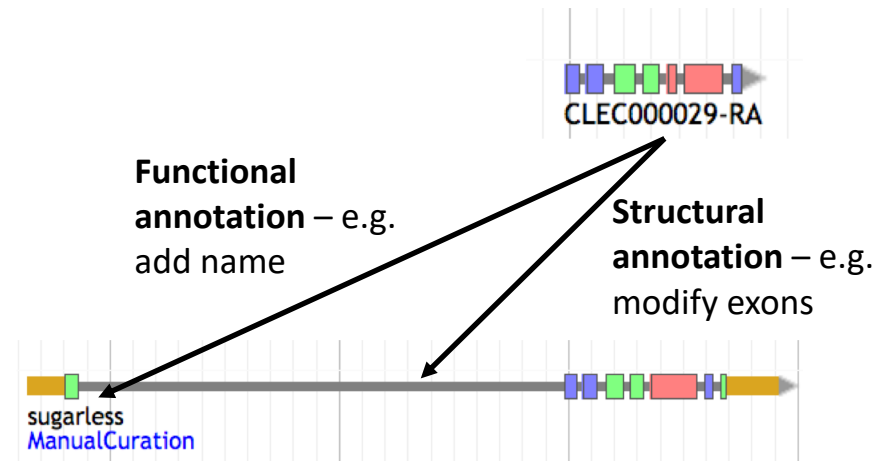
Other resources

- Monica Munoz-Torres from the Apollo group has a number of comprehensive tutorials:
 - <https://www.slideshare.net/MonicaMunozTorres/presentations>
 - I recommend these slides if you need more background:
 - <https://www.slideshare.net/MonicaMunozTorres/apollo-workshop-at-ksu-2015>
 - Note - there are two versions of Apollo. Some organisms at the i5k Workspace still use the older version with a slightly different interface
 - If you are new to Apollo, or need a refresher, I **highly recommend** that you review one of her presentations
- The official Apollo annotation guide:
 - <http://genomearchitect.org/users-guide/>
- Other manual curation tutorials:
 - <https://i5k.nal.usda.gov/manual-curation-example>
 - <http://genomecuration.github.io/genometrain/d-feature-curation-crossing/>

Manual annotation general overview

What is manual annotation?

- Manual review and improvement of an existing gene prediction
- Draw on external evidence (e.g. RNA-Seq, cDNA, genes from other species) to improve a computationally predicted gene model



Why manually annotate?

- “Incorrect annotations poison every experiment that makes use of them ... Worse still, the poison spreads because incorrect annotations from one organism are often unknowingly used by other projects to help annotate their own genomes.”
 - Yandell and Ence 2012, doi:10.1038/nrg3174
- Link gene models to existing literature and ontologies, providing richer data

General process of manual annotation

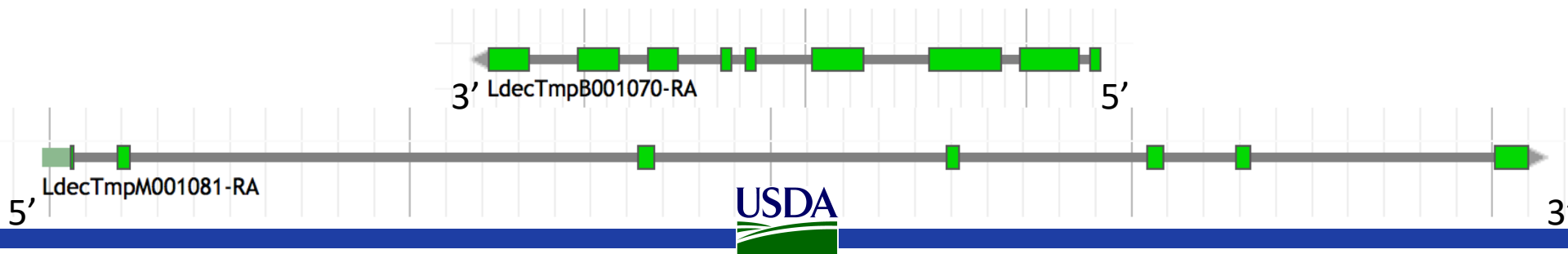
1. Select a chromosomal region of interest (e.g. scaffold)
 1. E.g. find sequence of interest from one or several other species, and align against proteins or genome sequence from your species
2. Select appropriate evidence (tracks in Apollo, or your own files)
3. Determine whether a feature in your evidence provides a reasonable starting gene model
 1. If yes: select and drag the feature to the 'user-created annotations' area, creating an initial gene model. If necessary use editing functions to adjust the model.
 2. If not – get in touch with us!
4. Edit model if necessary
5. Check your edited gene model for integrity and accuracy by comparing it with available homologs
 1. Verify that the gene model is the best representation of the underlying biology
6. Repeat steps 1 through 5 as needed to refine model
7. Add annotation details in the “Information Editor”
 1. Name, symbol, other comments

Adapted from <https://www.slideshare.net/MonicaMunozTorres/apollo-workshop-at-ksu-2015>

Manual annotation: i5k Workspace tools

First, some conventions

- HSP – High scoring pair in BLAST/BLAT alignments
 - The ‘Hits’ in an alignment result set
 - A subsection of a pair of sequences with sufficient score
 - HSPs can change based on the alignment parameters
- Five prime end and three prime end
 - Based on direction of transcription
 - Initiation site is at the five prime end
 - Stop codon is at the three prime end
- In the genome browser, arrowheads indicate direction



JBrowse and Apollo2

File: Add your own files

View: Change coloring scheme

Tools: Search using BLAT

Locate where you are on the scaffold

Search for a gene or location

Apollo2 Track selector

Revert to 'old' track selector

Log out

Zoom in/out

User-created annotations track

Find information about tracks

JBrowse is a web- based genome browser

- Visualize features that are mapped to a genome
- These features are displayed as tracks
- Many different types of data may be displayed

Apollo adds editing functions to JBrowse

- Manual gene curation
- Changes automatically saved back to server
- Edits are visible to other annotators in real-time
- Editing history is tracked

Apollo2 – Annotations Panel

The screenshot shows the Apollo2 web interface. The main panel displays a genomic track for *Onthophagus taurus* Scaffold5. A user-created annotation, 'test mRNA', is highlighted in yellow. Below the track, various genomic features are visible, including gaps in assembly, OTAU models, and embryonic reads. The right-hand panel, titled 'Annotations', shows a list of annotations. The 'test mRNA' annotation is selected, and its details are displayed in the 'Details' tab. The 'Coding' tab is also visible, showing a table of exons and CDS regions. Callouts with arrows point to various features: the 'Annotations' panel, the 'Filter annotations' section, the 'View annotation overview' table, the 'Click on arrow to jump to annotation' button, the 'View functional annotation details' section, and the 'Can modify individual features via 'Coding' Panel' section.

Annotations panel

Filter annotations

View annotation overview

Click on arrow to jump to annotation

View functional annotation details

Can modify individual features via 'Coding' Panel

Name	Seq	Type	Length	Updated
OtauTmpA001262-RA	Scaffold1	gene	11,946	May 28, 2018
test gene	Scaffold5	gene	28,376	Jun 04, 2018
test mRNA		mRNA	28,376	Jun 04, 2018

Type	Start	Length
exon	2,817,179	677
exon	2,806,566	226
exon	2,832,890	2,052
CDS	2,806,608	27,094
exon	2,824,806	152
exon	2,832,682	150
exon	2,830,616	128

Apollo2 – Ref Sequence Panel

The screenshot displays the Onthophagus taurus genome browser interface. The top panel shows the genome assembly with a scale from 0 to 3,000,000. The middle panel displays user-created annotations, including scaffolds and genes. The bottom panel shows O. taurus embryonic reads and junctions. On the right, a 'Ref Sequences' panel is visible, showing a list of scaffolds with their lengths and annotations. The list includes scaffolds such as Scaffold5, Scaffold54, Scaffold527, Scaffold540, Scaffold58, Scaffold584, Scaffold500, Scaffold56, and Scaffold52. The 'Ref Sequences' panel also includes a search bar, a length filter, and an export button.

Annotations:

- User-created Annotations: Scaffold1:451216..464455 (13.24 Kb), OtatmpA001262-RA-00001
- OTAU_v0.5.3-Models: OtatmpM001261-RA, OtatmpA001262-RA
- O. taurus embryonic reads: JUNCO0000280, JUNCO0000284, JUNCO0000286, JUNCO0000290
- O. taurus embryonic junctions: JUNCO0000280, JUNCO0000284, JUNCO0000286, JUNCO0000290

Ref Sequences:

Name	Length	Annotations
Scaffold5	4,952,630	1
Scaffold54	1,151,439	0
Scaffold527	1,104,991	0
Scaffold540	986,169	0
Scaffold58	897,936	0
Scaffold584	779,470	0
Scaffold500	685,874	0
Scaffold56	465,482	0
Scaffold52	406,915	0

View Tools Help demo@demo.org > Onthophagus Scaffold1

1,500

16,464

Onthophagus taurus

Export 1 sequence(s) from Onthophagus taurus as GFF3

GFF3 GFF3 with FASTA

Export Cancel

Scaffolds

All Selected (1)

1-50 of 1,111

Name	Length	Annotations
Scaffold5	4,952,630	
Scaffold54	1,151,439	
Scaffold527	1,104,991	
Scaffold540	986,169	

i5k Workspace BLAST: one way to access Apollo

The screenshot shows the i5k Workspace BLAST interface. At the top is a navigation bar with the i5k@NAL logo and links for Tools, About Us, and Contact. The main content area is divided into sections: BLAST Databases, Query Sequence, and Program. The BLAST Databases section has a list of organisms on the left and a list of database types on the right. The Query Sequence section has a text input area and a file upload button. The Program section has radio buttons for different BLAST programs. Annotations with arrows point to specific elements: 'Select organism' points to the organism list; 'Paste or upload query sequence(s)' points to the query sequence input; 'Program is automatically selected' points to the 'tblastn' radio button; 'Select organism-specific database' points to the 'Genome Assembly' database; and 'BLAST against the genome assembly to view HSPs in Jbrowse' points to the 'Genome Assembly' database.

Select organism

Paste or upload query sequence(s)

Program is automatically selected

Select organism-specific database

BLAST against the genome assembly to view HSPs in Jbrowse

BLAST Databases

Organisms

- ☐ *Drosophila takahashii*
- ☐ *Dufourea novaeangliae*
- ☐ *Ephemera danica*
- ☒ *Eurytemora affinis*
- ☐ *Fopius arisanus*
- ☐ *Frankliniella occidentalis*
- ☐ *Gerris buenoi*
- ☐ *Habropoda laboriosa*
- ☐ *Halyomorpha halys*
- ☐ *Homalodisca vitripennis*
- ☐ *Hyaella azteca*
- ☐ *Ladona fulva*
- ☐ *Lasioglossum albipes*

Eurytemora affinis

Nucleotide

- ☒ Genome Assembly - Eaff_11172013.genome_new_ids.fa
- ☐ Transcript - EAFF_new_ids.fna

Peptide

- ☐ Protein - EAFF_new_ids.faa

Query Sequence

Your sequence is detected as peptide:

>FBpp0070332
MDNCDQDASFRLSHIKEEVKPDISQLNDSNN
SSFSPKAESPVPFMQAMSMVHVLPGSNSASS
NNNSAGDAQMAQAPNSAG
GSAAAQVQYPPNHPLSGSKHLCSICGDRA
SGKHVGVYSCGCKGFFKRTVRKDLTYACRE

Or load it from disk

No file selected.

Program

☐ tblastn ☒ tblastn ☐ tblastx ☐ tblastp ☐ tblastx

tblastn - Peptide vs. Translated Nucleotide

URL: <https://i5k.nal.usda.gov/webapp/blast/>

i5k Workspace BLAST: one way to access Apollo

Query Coverage Graph - FBpp0070332, BLAST Hits 1-9

Subject Coverage Graph - gnl|Eurytemora_affinis|euraff_Scaffold427, BLAST Hits 1-9

Showing 1 to 9 of 9 entries (filtered from 55 total entries)

blastdb	qseqid	sseqid	pid	length	mismatch	gapopen	qstart
euraff	FBpp0070332	Scaffold427	36.36	77	49	0	419
euraff	FBpp0070332	Scaffold427	26.67	165	83	4	262
euraff	Eaff_11172013.genome_new_ids.fa	EaffTmpM006514-RA-00001	59.21	76	31	0	103
euraff	FBpp0070332	Scaffold4229	56.52	92	37	1	98
euraff	FBpp0070332	Scaffold200	57.14	91	36	1	99
euraff	FBpp0070332	Scaffold12	58.57	87	39	2	104
euraff	FBpp0070332	Scaffold12	58.57	87	39	2	104
euraff	FBpp0070332	Scaffold13	85.71	35	5	0	91
euraff	FBpp0070332	Scaffold200	58.62	81	38	1	101

Click on blue blastdb icon next to your favorite HSP

BLAST result page with 4 panels

BLAST Report

FASTA

Score = 86.3 bits (212), Expect = 2e-16, Method: Compositional matrix adjust.

Identities = 45/76 (59%), Positives = 61/76 (80%), Gaps = 0/76 (0%)

Query 103

Subject 255352

Score = 62.4 bits (155), Expect = 3e-13, Method: Compositional matrix adjust.

Identities = 33/76 (43%), Positives = 51/76 (67%), Gaps = 0/76 (0%)

Query 414

Subject 251038

Query 474

Eurytemora affinis - training

File View Tools Help

Scaffold33

Scaffold33:1487368..1488330 (964 b)

Annotations

0. Reference Assembly

1/3

BLAST Results

Gaps in assembly

GC Content

BCM_v0.5.3/1. Gene Sets/Primary Gene Sets: Protein Coding

1/1

EAFV_v0.5.3-Models

Blast results are displayed in Apollo

HMMER and Clustal

- Use HMMER to detect remote protein homologs
- <https://i5k.nal.usda.gov/webapp/hmmer/>
- Use Clustal to perform multiple sequence alignments
- <https://i5k.nal.usda.gov/webapp/clustal/>

Tips and Tricks

- The i5k Workspace BLAST results persist for one week
 - You can bookmark and share searches
 - BLAST HSPs are ‘draggable’ and can be used in annotations
- Jbrowse/Apollo URLs can be shared
 - Allow you to share the exact view (including active tracks) with others
 - Great for troubleshooting with collaborators
- In Apollo “walk” feature boundaries
 - Square brackets walk exon boundaries: [and]
 - Curly brackets walk gene boundaries: { and }
- In Apollo, you can pin tracks to the top
- If you know the name or ID of the gene that you’d like to annotate, you can paste it into the search box in Apollo to navigate to it

Manual annotation example: preparation

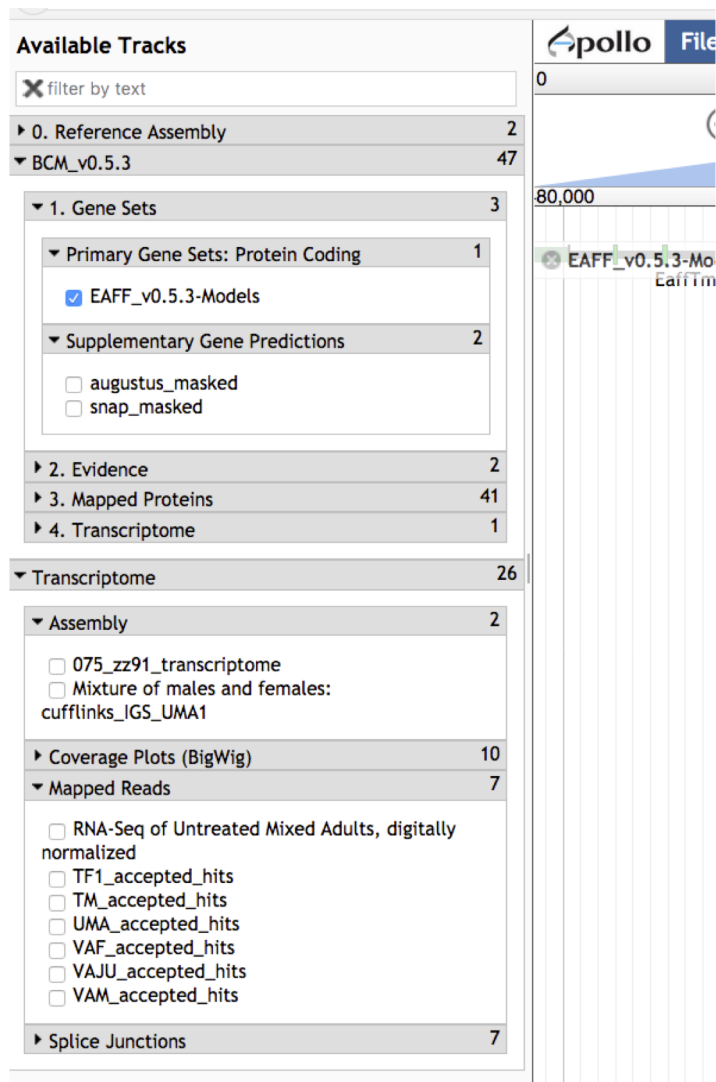
Annotation Example

- Phosphoenolpyruvate carboxykinase (pepck) in the copepod *Eurytemora affinis*
- Pepck catalyzes the conversion of oxaloacetate (OAA) to phosphoenolpyruvate (PEP).
- More information about the copepod:
https://i5k.nal.usda.gov/Eurytemora_affinis
- Apollo URL (for training only):
<https://apollo.nal.usda.gov/apollo/3068161/jbrowse/index.html>
 - Login credentials: demo/demo

Notes on *E. affinis* genome/browser

- Big advantage for annotation: lots of RNA-Seq and transcriptome data are available to use as contributing evidence for your gene models
 - Includes strand-specific RNA-Seq
- Disadvantage: No close reference genomes, so it may be harder to find homologs for your genes of interest to inform your annotations.

Available tracks for *E. affinis*



The screenshot displays the Apollo genome browser interface. On the left, a sidebar titled 'Available Tracks' lists various genomic data categories and their track counts. The categories include Reference Assembly (2 tracks), Gene Sets (3 tracks), Evidence (2 tracks), Mapped Proteins (41 tracks), Transcriptome (26 tracks), and Splice Junctions (7 tracks). The 'Gene Sets' section is expanded, showing 'Primary Gene Sets: Protein Coding' (1 track) with 'EAFF_v0.5.3-Models' selected, and 'Supplementary Gene Predictions' (2 tracks) with 'augustus_masked' and 'snap_masked' options. The 'Transcriptome' section is also expanded, showing 'Assembly' (2 tracks) with '075_zz91_transcriptome' and 'Mixture of males and females: cufflinks_IGS_UMA1' options, 'Coverage Plots (BigWig)' (10 tracks), 'Mapped Reads' (7 tracks) with various RNA-Seq libraries, and 'Splice Junctions' (7 tracks). The main panel on the right shows a genomic track view with a blue line representing the 'EAFF_v0.5.3-Models' track and a green line representing the 'EAFF_v0.5.3-Mo' track. The Apollo logo and a 'File' button are visible at the top of the main panel.

Category	Track Count
0. Reference Assembly	2
BCM_v0.5.3	47
1. Gene Sets	3
Primary Gene Sets: Protein Coding	1
EAFF_v0.5.3-Models	
Supplementary Gene Predictions	2
augustus_masked	
snap_masked	
2. Evidence	2
3. Mapped Proteins	41
4. Transcriptome	1
Transcriptome	26
Assembly	2
075_zz91_transcriptome	
Mixture of males and females: cufflinks_IGS_UMA1	
Coverage Plots (BigWig)	10
Mapped Reads	7
RNA-Seq of Untreated Mixed Adults, digitally normalized	
TF1_accepted_hits	
TM_accepted_hits	
UMA_accepted_hits	
VAF_accepted_hits	
VAJU_accepted_hits	
VAM_accepted_hits	
Splice Junctions	7

- Baylor Maker annotations:
 - Primary Gene Set:
 - EAFF_v0.5.3-Models
 - Other tracks that were used to generate the primary gene set
- Transcriptome/RNA-Seq
 - Transcriptome assemblies
 - Coverage plots, Mapped RNA-Seq data, Splice junctions
 - Some of the RNA-Seq libraries are stranded

Choosing reference proteins: *D. melanogaster* pepck in UniProt

UniProtKB - P20007 (PCKG_DROME)

Display

- Entry
- Publications
- Feature viewer
- Feature table
- All None
- Function

BLAST Align Format Add to basket History

Protein | Phosphoenolpyruvate carboxykinase [GTP]
Gene | Pepck
Organism | *Drosophila melanogaster* (Fruit fly)
Status | Reviewed - Annotation score: ●●●○○○ - Experimental evidence at transcript levelⁱ

Annotation score is a heuristic for annotation quality

Organism-specific databases

FlyBaseⁱ FBgn0003067. Pepck.

Subcellular locationⁱ

Flybase is another great resource

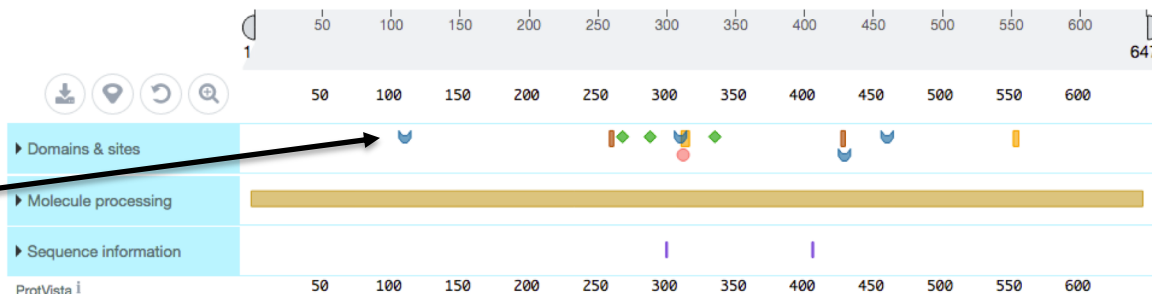
UniProtKB - P20007 (PCKG_DROME)

Display

- Entry
- Publications
- Feature viewer
- Feature table

BLAST Align Format Add to basket History

Feeds



Feature viewer gives graphical view of domains and sites

Catalyzes the conversion of oxaloacetate (OAA) to phosphoenolpyruvate (PEP).

Source: <http://www.uniprot.org/uniprot/P20007>

Choosing reference proteins: *Daphnia pulex* Pepck

- GenBank record:

<https://www.ncbi.nlm.nih.gov/protein/EFX80236.1>

Lynch, M., Boore, J.L. and Grigoriev, I.V.

CONSRTM US DOE Joint Genome Institute (JGI-PGF)

TITLE Direct Submission

JOURNAL Submitted (02-FEB-2011) US DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598-1698, USA

COMMENT Method: conceptual translation.

FEATURES Location/Qualifiers
source 1..652

← Treat with caution!!!

Phosphoenolpyruvate carboxykinase,

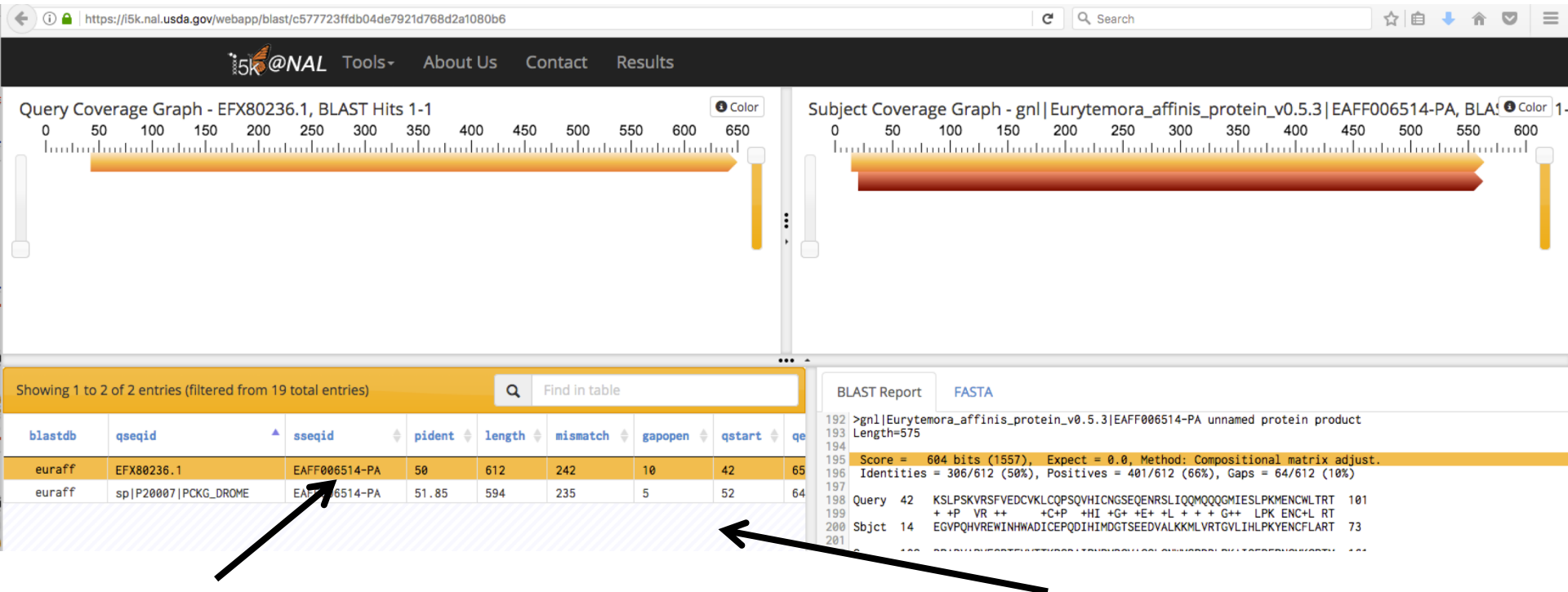
(daphnia Phosphoenolpyruvate carboxykinase)

(daphnia Phosphoenolpyruvate carboxykinase)

Manual annotation live example

BLAST dmel, dpul proteins against *E. affinis* proteins

<https://i5k.nal.usda.gov/training/webapp/blast/>



Copy the protein 'base name'
EAF006514 for searching in Apollo

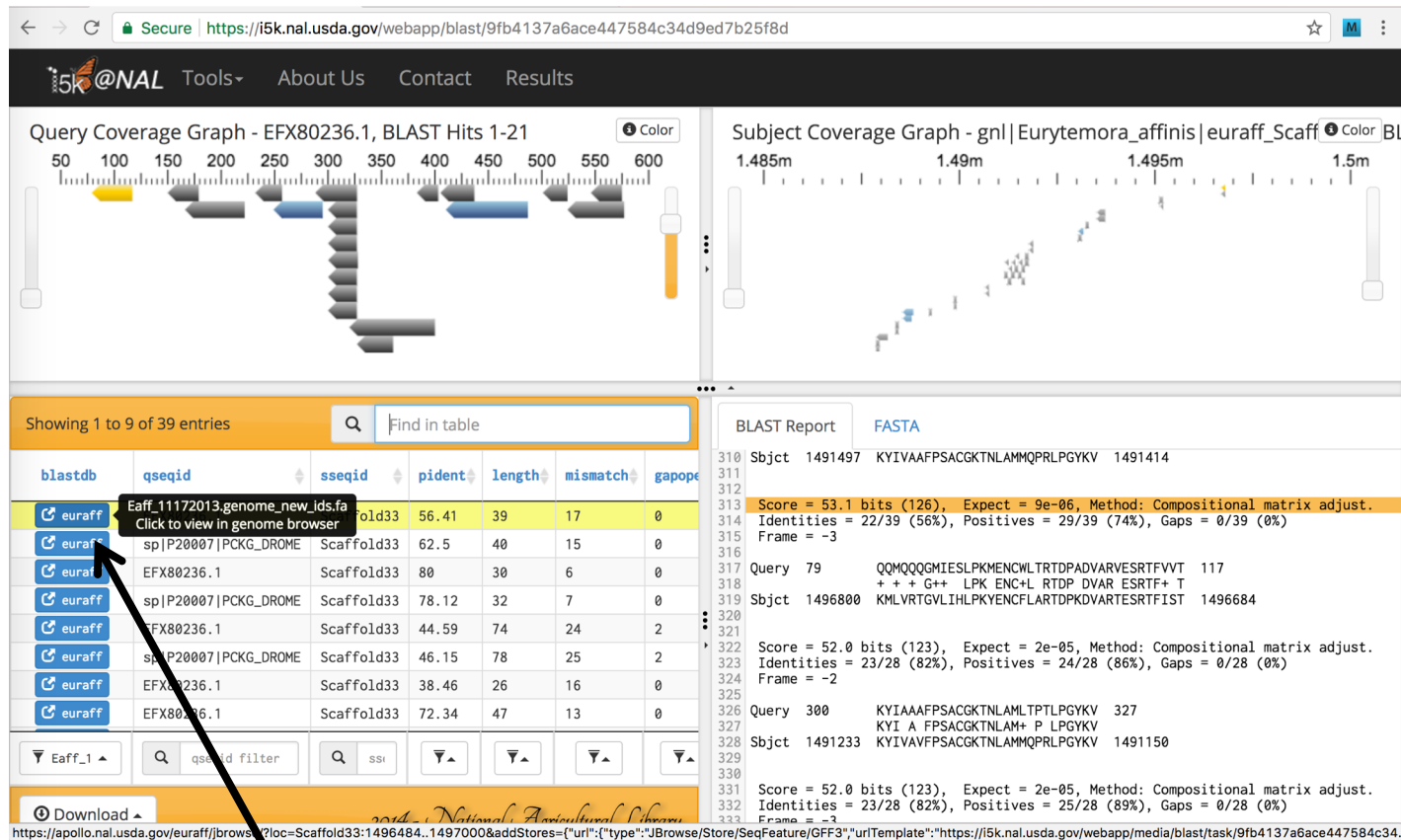
Results are filtered by e-value; only
one protein in the *E. affinis* dataset has
a significant match

Modify *E. affinis* model sequence in Apollo

- Go to Apollo URL:
<https://apollo.nal.usda.gov/apollo/3068161/jbrowse/index.html>
 - Find mRNA of EAFF006514-PA in genome browser by pasting EAFF006514 into search box, selecting EAFF006514-RA
- Log in to Apollo
- Drag EAFF006514-RA into the yellow annotation track
- Check available evidence for model

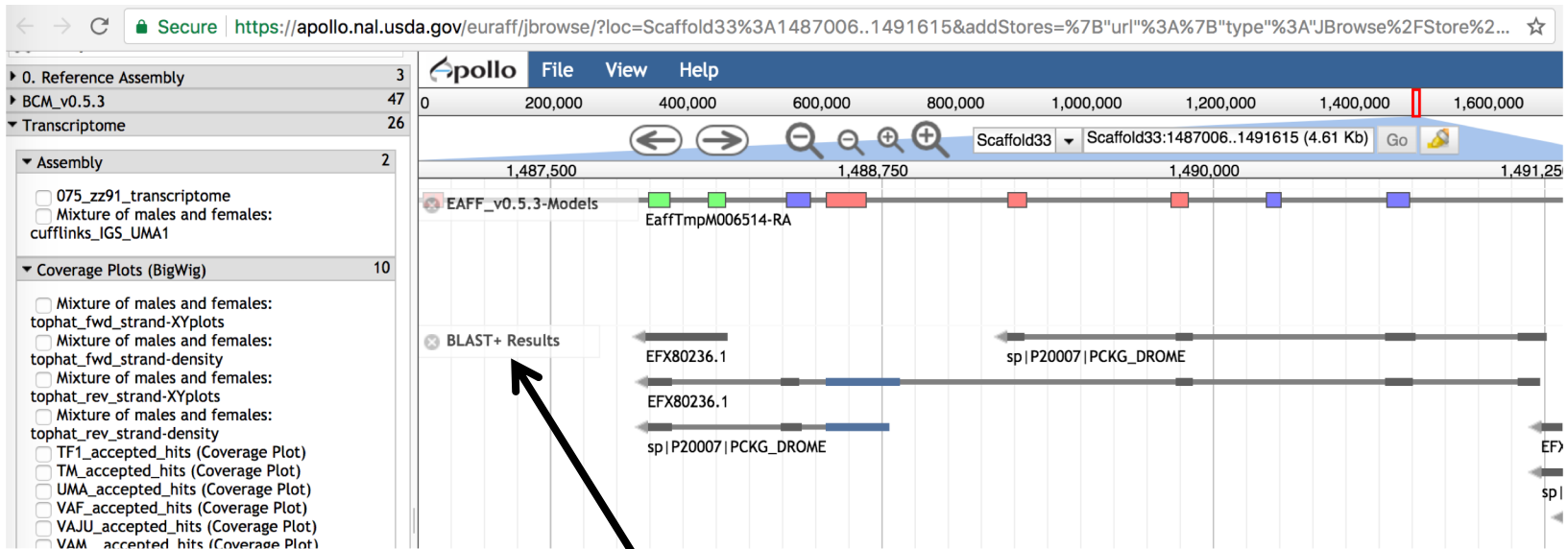
Another approach: BLAST against the genome

<https://i5k.nal.usda.gov/training/webapp/blast/>



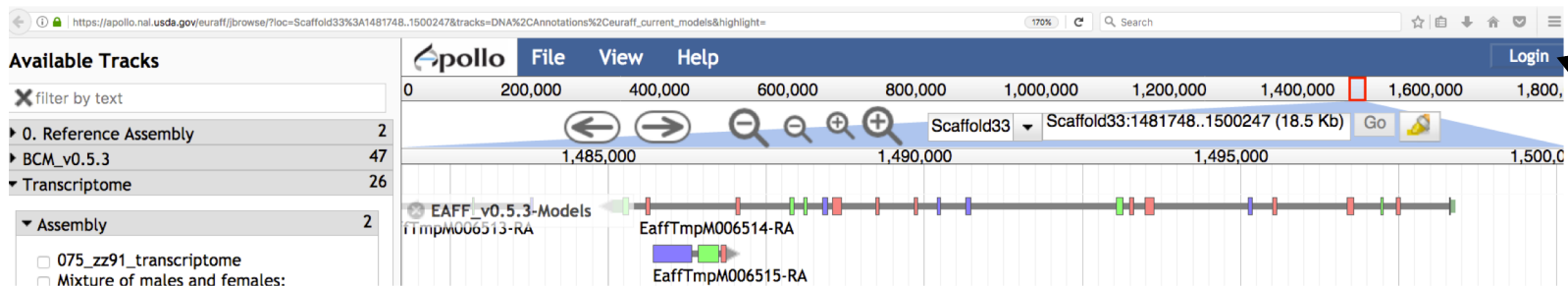
Click on blue blastdb button next to your favorite HSP to view it in JBrowse

Another approach: BLAST against the genome

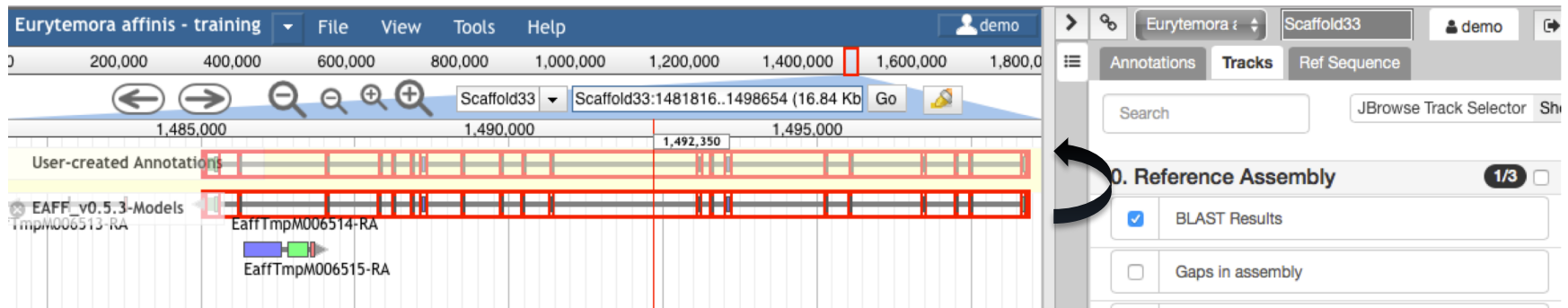


BLAST results are displayed as glyphs in browser;
can be used as annotation starting points if the
alignment is high quality

Create annotation in user-created annotations track



Log in with
your
Apollo
credentials

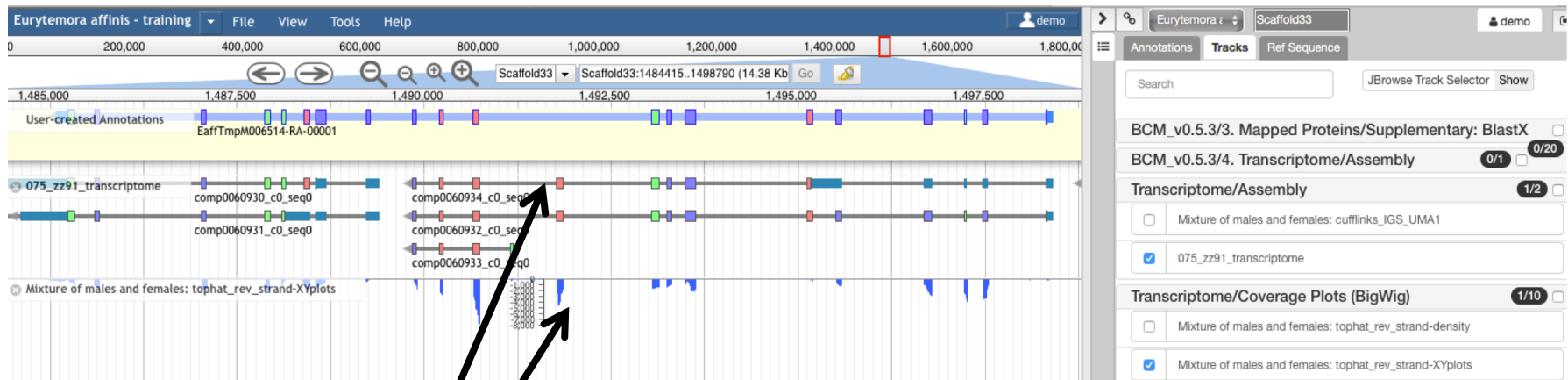


Drag model EaffTmpM006514-
RA to User-created Annotations
track

Modify *E. affinis* model sequence in Apollo

- Questions:
 - What evidence do you choose to check the integrity of the model?
 - Do you need additional evidence?
 - How do you evaluate whether the protein sequence is as complete as it can be?
 - Should you add/modify UTRs?

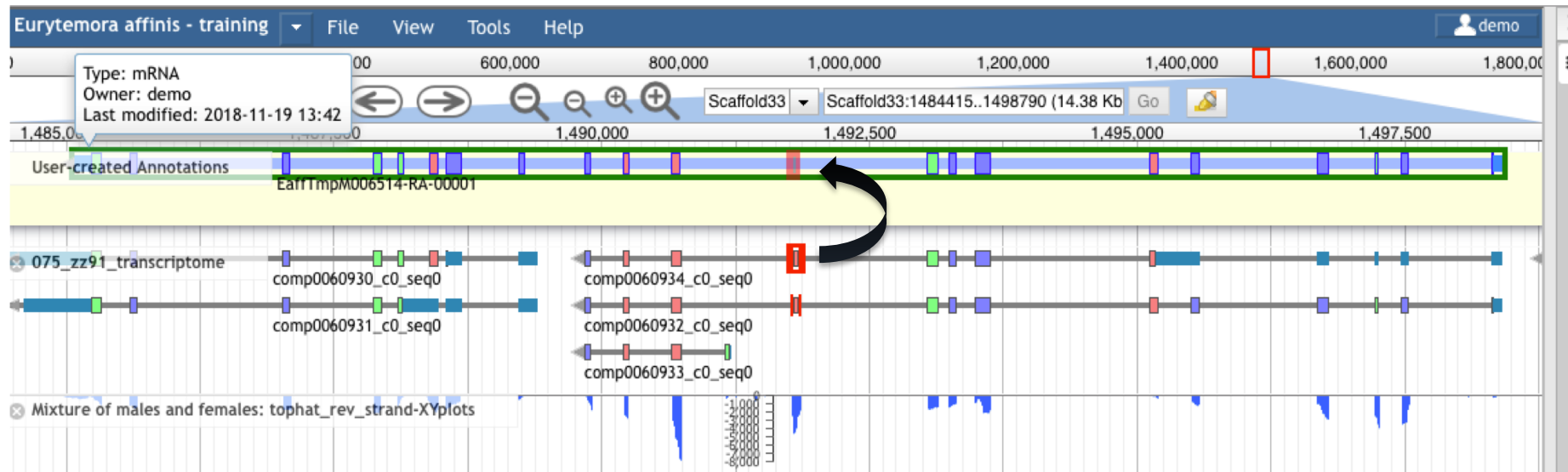
View available evidence



RNA-Seq and transcriptome tracks suggest that one exon is missing

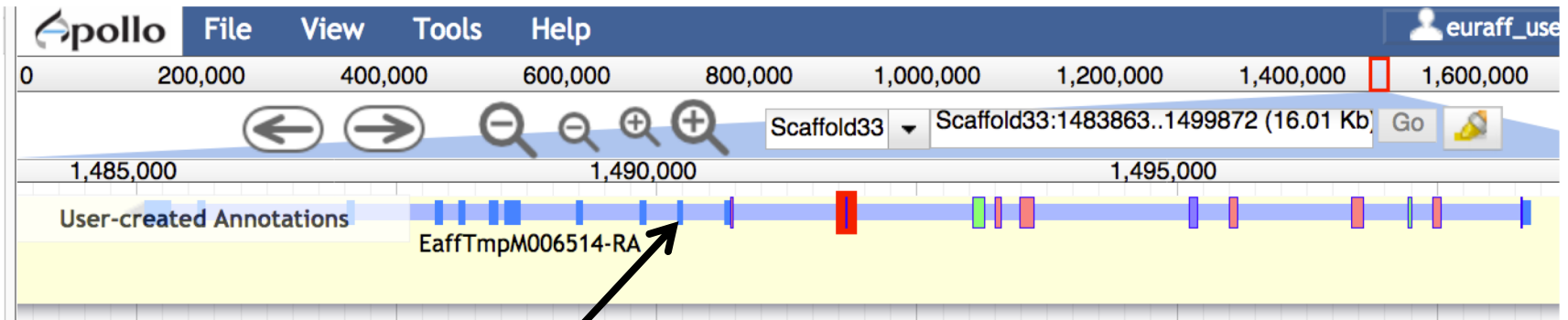
Model is on the reverse strand, so we can take advantage of the stranded RNA-Seq available for this species

Add an exon to the model

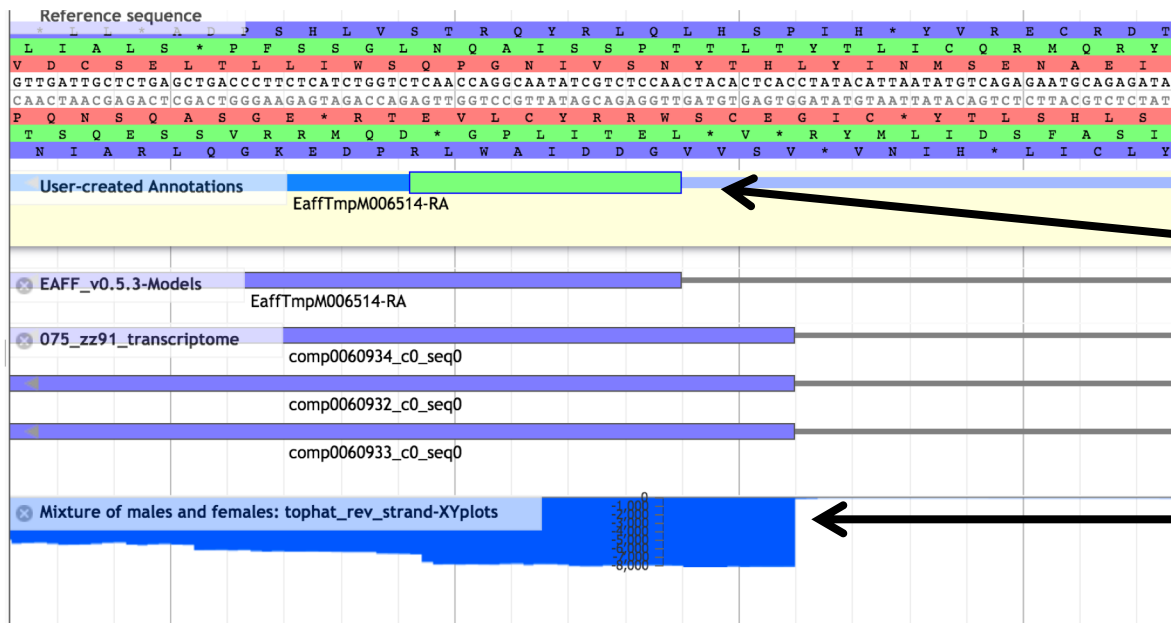


Drag exon from
transcriptome track
into new gene model

Adjust exon boundary



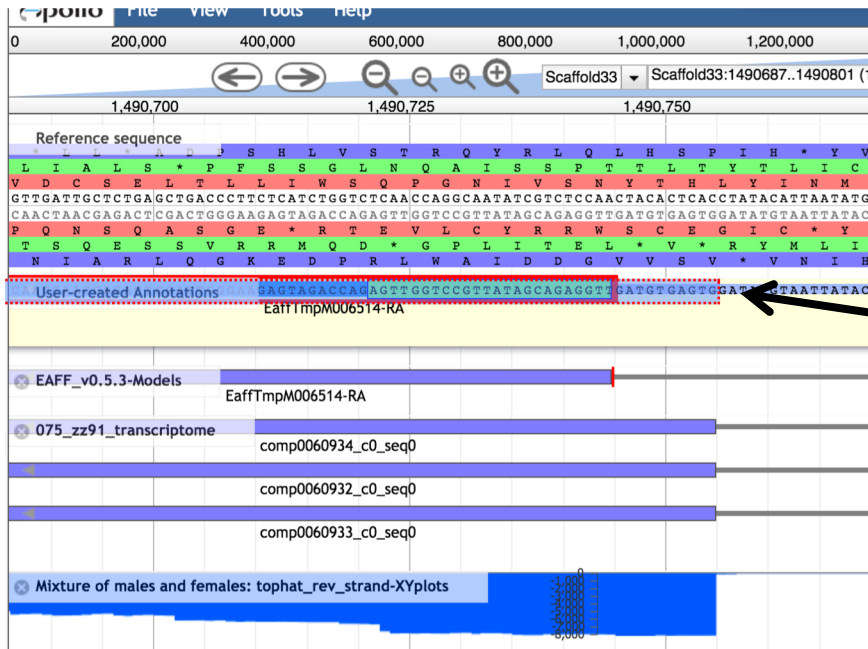
CDS sequence is now UTR – zoom in to investigate



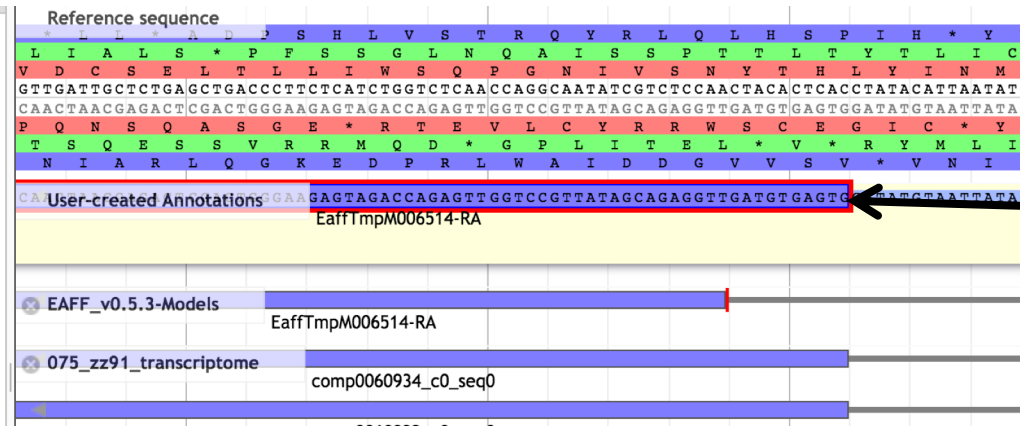
CDS frame has changed from purple to green—we need to fix this

RNA-Seq suggests we need to adjust exon boundary

Adjust exon boundary



Drag exon boundary to match RNA-Seq and transcriptome tracks



Fixed both reading frame and exon boundary

Evaluate new protein sequence

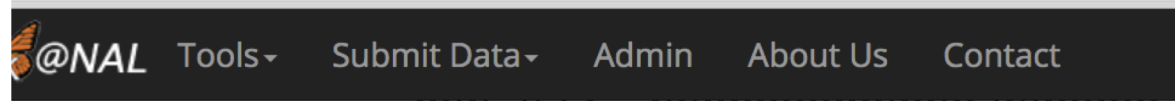
- Blast modified EAFF006514-PA sequence to NCBI's nr database
 - Make sure it doesn't match a potential contaminant
 - Get an idea whether you have the right sequence
 - Blastp home:
 - https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome
- Once contamination is ruled out, it's better to align your sequence against a smaller set of high-quality proteins
- If you notice that parts of the protein are missing, check the 'Gaps in assembly' track in the browser

Evaluate new protein sequence

- Get *E. affinis* pepck protein sequence from old model and new model
- Align new and old sequence to dmel and dmag protein sequences
 - Clustal (<https://i5k.nal.usda.gov/webapp/clustal/>)
 - Can also use NCBI Blast
- Check alignment extent, %ID

Clustal Results

:/f5k.nal.usda.gov/webapp/clustal/105850a3594e4234a21b07d93cbbd71



euraff_old_pepck
euraff_new_pepck
sp|P20007|PCKG_DROME
EFX80236.1

```
IS-----VGDDIAWLRPDEKQQLRAI
ISGITNSQGEKKYIVAAFPSCGKTNLAMMQPRLP
ILGITDPKGEKKYITAAFPSCGKTNLAMNPSLANYKVECVGDDIAWMKFD SQVLRAI
ILGITNPQGQKKYIAAAPPSCGKTNLAMLTPTLPGYKVECVGDDIAWMHFDKEGLRAI
*                               *****: :*.:* ****
```

New exon added

euraff_old_pepck
euraff_new_pepck
sp|P20007|PCKG_DROME
EFX80236.1

```
NPENGFFGVAPGTSYTSNPVA-----MQSIFKDTIFSNVAMTDDGGVWVEGMGDKPK
NPENGFFGVAPGTSYTSNPVA-----MQSIFKDTIFSNVAMTDDGGVWVEGMGDKPK
NPENGFFGVAPGTSMETNPPIA-----MNTVFKNITFTNVASTSDGGVFWEGMESSLA
NPENGFFGVAPGTNYATNPACYNFFLYAMLTIQKNTIFTNVAKTSDGGVFWEGLEKEV-
*****: :* * * : : * :*:*: * :*:*:*: :.
```

euraff_old_pepck
euraff_new_pepck
sp|P20007|PCKG_DROME
EFX80236.1

```
ERSSCIDWKGK-PWRPTSSNPAHPNSRFCTPLLNCVLPDESAEDPAGVPFAAILFGGRR
ERSSCIDWKGK-PWRPTSSNPAHPNSRFCTPLLNCVLPDESAEDPAGVPFAAILFGGRR
PNVQITDWLGK-PWTKDSGKPAHPNSRFCTPAAQCPIIDEAWEDPAGVPISAMLFGGRR
TGVDTISWLGDNWTKSSGKPAHPNSRFCCAPASQCPIIDPLWESPEGVPIIDAILFGGRR
. . * * * * :*:*:*:*: * :*:*: * :* * * :*:*:*
```

euraff_old_pepck
euraff_new_pepck
sp|P20007|PCKG_DROME
EFX80236.1

```
PSGVPLVYQAISWEHGVFMGACVKSEATAAAEFKGQIMHDPFMRPFFG-----HW
PSGVPLVYQAISWEHGVFMGACVKSEATAAAEFKGQIMHDPFMRPFFG-----HW
PAGVPLIYEARDWTHGVFI GAAMRSEATAAAEHKGKVMHDPFMRPFFGYNFGDYVAHW
PRGVPLVYEALNWKHGVFVGASVSEATAAAEHKGRS IMHDPFMRPFFGYNAGNYLGHW
* ****:*: * . * ****:*. : : *****:*. : *****:***** **
```

Another exon might be missing (we're not going to handle this today)

Using the Information Editor

The screenshot shows the 'Information Editor' window with a genomic track at the top (400,000 to 1,400,000) and a 'Select mRNA' dropdown set to 'EaffTmpM006514-RA-00001'. The window is split into two panes: 'gene' on the left and 'mRNA' on the right. Both panes have identical form fields: Name, Symbol, Description, Created (2018-11-19), Last modified (2018-11-19), Status (radio buttons for Approved and Delete), and a table for DBXRefs with columns DB and Accession. The 'mRNA' pane has the Name field filled with 'Phosphoenolpyruvate carboxykinase' and the Symbol field filled with 'Pepck'.

gene	
Name	
Symbol	
Description	
Created	2018-11-19
Last modified	2018-11-19
Status <input type="radio"/> Approved <input type="radio"/> Delete	
DBXRefs	
DB	Accession

mRNA	
Name	Phosphoenolpyruvate carboxykinase
Symbol	Pepck
Description	
Created	2018-11-19
Last modified	2018-11-19
Status <input checked="" type="radio"/> Approved <input type="radio"/> Delete	
DBXRefs	
DB	Accession

Use the
mRNA/transcript
side of the IE

Review our naming guidelines
before naming:

<https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines>

Using the Information Editor

- Select the model in Apollo, then right-click, and select 'Edit Information' from the drop-down menu
 - Use the 'mRNA' section
 - Name: We recommend the INSDC naming guidelines:
 - <http://www.uniprot.org/docs/nameprot>
 - If a naming convention exists, use it (e.g. for gene families)
 - Name should be unique and attributed to all orthologs (as far as possible)
 - Use name from an orthologous protein if you are sure that your gene model is an ortholog.
 - Document your justification for the name in the Comments field (e.g. "88% sequence similarity via blastp to D. melanogaster pepck P20007")
 - Comments – Document what changes you performed, and your justification for the name. These notes will be visible in the OGS, so make sure that others understand them

Checklist for accuracy and integrity

- Check start, stop and exon boundaries (splice sites)
 - Try to fix non-canonical splice sites if possible
- Check if you can annotate UTRs (e.g. using RNA-Seq data)
- Check for gaps in the genome
- If you change the genome sequence, add a justification comment to the corresponding gene model
- Use BLAST or a multiple sequence aligner
 - To look at completeness of model
 - To verify the appropriateness of the gene name
- In the Information editor **mRNA** field
 - Update the Name if appropriate
 - Add comments that describe
 - your evidence for the annotation
 - Modifications that you made to the gene model

cf. <https://www.slideshare.net/MonicaMunozTorres/editing-functionality-apollo-workshop>

What happens to my annotation when I'm done?

- This depends on the genome project that you're working on.
- If the genome coordinator has asked us to generate an OGS (Official Gene Set), we will do so
 - We are still working on this process, so if you ask us to do this, 1) it will take some time, and 2) we will probably ask you for co-authorship if you publish a paper on the OGS.
 - We are working on a pipeline to submit Official Gene Sets to GenBank, where they will be archived/accessioned
- Otherwise, don't assume that your annotation will be archived.
 - If you need it to be, get in touch with us and we'll figure out what to do.
- Get in touch with us and the genome project coordinator if you're not sure about the status of a genome project.
- <https://i5k.nal.usda.gov/data-management-policy>

I5k Workspace ‘Etiquette’

1. Use Apollo to improve a gene model in an i5k Workspace assembly.
 1. If you just want to practice – use one of our training instances.
 1. <https://i5k.nal.usda.gov/jbrowseapollo-training>
 2. If you just want to view the data – you probably can get what you want without using Apollo. All of the data that we host is public.
2. Your annotation work is a community effort.
 1. If you notice that someone else is working on your model of choice, get in touch with them (or us) and collaborate – don’t make a 2nd model or delete the other model.
 2. Keep in mind that your work may be used by the scientific community once you’re done.
3. If you publish any of your work generated in the i5k workspace:
 1. Get in touch with the genome contact first (you can find the contact info on the organism page; <https://i5k.nal.usda.gov/species>);
 2. Please cite the i5k Workspace paper! This helps us continue to exist.
 1. <https://doi.org/10.1093/nar/gku983>

Thank you!

The NAL Team

- Chris Childers
- Gary Moore
- Susan McCarthy
- Min-Chen Hsu
- Chun-Hung Lin
- Chia-Tung Wu

I5k Workspace alumni

- Yu-yu Lin
- Chaitanya Gutta
- Li-Mei Chiang
- Yi Hsiao
- Chien-Yueh Lee
- Han Lin
- Jun-Wei Lin
- Vijaya Tsavatapalli
- Mei-Ju Chen
- Chao-I Tuan

- i5k Coordinating Committee
- i5k Pilot Project
- Apollo & JBrowse Development Teams
- GMOD/Tripal community
- All of our users and contributors!

Contact us:

- <https://i5k.nal.usda.gov/contact>
- i5k@ars.usda.gov
- Monica.Poelchau@usda.gov
- Christopher.Childers@usda.gov

